Notes on Probability and Statistics

Andrew Zou Li

Last Updated March 7, 2022

Contents

1 Introduction to Probability

1.1 Uncertainty

There is always uncertainty in a description of a real world system. This could arise from imprecise measurements, incomplete models, and limited measurements. Probability and statistics allow us to describe uncertainty in a precise manner.

1.2 Set Notation

Definition: Sample Space
The set of all possible outcomes of a statistical experiment is called the sample space and is represented by the symbol S.

Example: $S = \{(x, y)|x^2 + y^2 \leq 4\}$
An **event** is a subset of a sample space.

Let there be two events A and B which are sets with elements/members.

- $A = B$ if they have the same elements
- $A \subset B$ if A is a subset of B, meaning each element in A is also in B
- $A \subseteq B$ if A is possible equal to B
- $a \in A$ means a is a member of set A
- $a \notin A$ means a is not a member of set A
- The complement of event A with respect to S is denoted A'
- The intersection of two events is all the elements common to both events, denoted $A \cap B$
- The union of two events is all members in A and B, denoted $A \cup B$
- If $A \cap B = \emptyset =$, they are mutually exclusive or disjoint
- $A\ B$ denotes the set of elements in A but not in B

There are some fundamental properties that apply for both union and intersection. Only union will be shown but swapping the operators will also be a valid expression.
`https://en.wikipedia.org/wiki/Algebra_of_sets`

- Commutative: $A \cup B = B \cup A$
- Associative: $(A \cup B) \cup C = A \cup (B \cup C)$
- Distributive: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- Identity: $A \cup \emptyset = A$ and $A \cap S = A$
- Complement: $A \cup A' = S$ and $AcapA' = \emptyset$

1.3 Counting

Combinatorics is too difficult. See Appendix A1 for examples and stuff.

Definition: Combination
$nCr = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

Definition: Permutation
$nPr = \frac{n!}{(n-r)!}$

1.4 Probability

Probability of A:

$$P(A) \in [0, 1]$$

Probability of Sample Space:

$$P(S) = 1$$

Probability of Mutually Exclusive Events:

$$P(A) \cup P(B) = P(A) + P(B)$$

Probability of Independent Events:

$$P(A, B) = P(A)P(B)$$

Probability of Two Events:

$$P(AB) = P(A) + P(B) - P(A \cap B)$$

Probability of Mutually Exclusive Events (probability of intersection is 0):

$$P(A) \cup P(B) = P(A) + P(B)$$

Probability of N Events (Intuitive with a Venn Diagram): [1]

$$P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Probability of Event and Complement

$$P(A) + P(A') = 1$$

Example 1
Cards Chance of getting 2 aces and 3 jacks in a regular card deck. Notes:

- 52 cards in a regular card deck
- 13 of each suite and colour
- 4 of each number/face
- 12 are face cards
- 40 are number cards

[1] Using distributivity of set operators,

$$\begin{aligned} P(ABC) &= P(A) + P(B \cup C)P(A \cap (B \cup C)) \\ &= P(A) + [P(B) + P(C) - P(B \cap C)] \\ &\quad - P(A \cap (B \cup C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) \\ &\quad - P((A \cap B \cup A \cap C)) \\ &= P(A) + P(B) + P(C) - P(A \cap B) \\ &\quad - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

$$P(2 \text{ aces}) = \binom{4}{2} = 6$$
$$P(3 \text{ jacks}) = \binom{4}{3} = 4$$
$$N_{event} = (6)(4) = 24$$
$$N_{possible} = \binom{52}{5}$$
$$P(event) = \frac{24}{\binom{52}{5}}$$

Example 2
Dartboard Chance of landing within an area A. Let the entire dartboard be S and thus area $A \subset S$.

$$P(A) = \frac{\int_A dx}{\int_S dx}$$

1.5 Conditional Probability

Definition: Conditional Probability
The probability of event B given event A is given by,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

If $P(A) > 0$

Product Rule:

$$P(A \cap B) = P(A)P(B|A)$$

Independence of two events is true iff $P(A|B) = P(A)$ or $P(B|A) = P(B)$. Thus, probability of independent events is given by,

$$P(A, B) = P(A)P(B)$$

Partition:
$B_1, \dots B_k$ is a partition of S if $B_i \cap B_j = \emptyset$ and $B_1 \cup B_2 \cup \dots B_k = S$

Theorem: Bayes' Rule
Suppose $P(A) > 0$, $P(B) > 0$ and $C_1, \dots C_k$ is a partition. Then,

$$P(B|A)P(A) = P(A|B)P(B)$$
$$\frac{P(B|A)}{P(B)} = \frac{P(A|B)}{P(A)}$$
$$P(B|A) = \frac{P(B)P(A|B)}{\sum_{i=1}^{k} P(C_i)P(A|C_i)}$$

Definition: Total Probability

Let A be an event and $B_1, \ldots B_k$ is a partition. Then,

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i-1}^{k} P(A|B_i)P(B_i)$$

Example 3
Machines B_1, B_2, B_3 make $30\%, 45\%, 25\%$ of the products at defective rates of $2\%, 3\%, 2\%$ respectively. Using the rule of total probability,

$$\begin{aligned} &P(\text{Defective}) \\ &\qquad = P(B_1)P(\text{Defective}|B_1) + P(B_2)P(\text{Defective}|B_2) \\ &\qquad\quad + P(B_3)P(\text{Defective}|B_3) \\ &\qquad = (0.3)(0.02) + (0.45)(0.03) + (0.25)(0.02) \\ &\qquad = 0.0245 \end{aligned}$$

We can reverse the direction of probability calculations using Bayes' rule. If a product is defective, then the probability it came from machine 3 is

$$\begin{aligned} P(B_3|D) &= \frac{P(B_3)P(D|B_3)}{P(D)} \\ &= \frac{(0.25)(0.02)}{0.0245} \\ &= \frac{10}{49} \end{aligned}$$

2 Random Variables

Definition: Random Variable (RV)
A random variable is a function that maps a non-negative real number with each element in the sample space.
It is denoted with a capital letter, e.g. X or Y.
Example: Let there be a subspace $E = AAB, ABA, BAB$. X is the random variable for outcome A and takes on the value $x = 2$ in the first two outcomes and $x = 1$ in the last outcome.

2.1 Discrete Probability Distributions

Probability Mass Function (PMF)

$f(x)$ is a PMF of the discrete RV X if

- $f(x) \geq 0$ for each outcome $X = x$
- $\Sigma_x f(x) = 1$
- $f(x) = P(X = x)$

Cumulative Distribution Function (CDF)

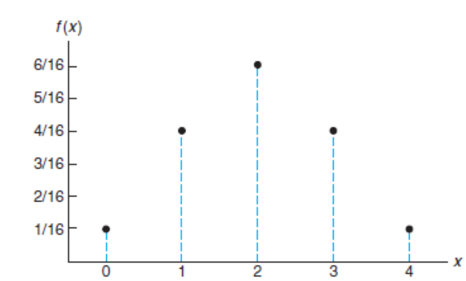

Figure 1: PMF as discrete points, can also be visualized as a histogram. Source: textbook

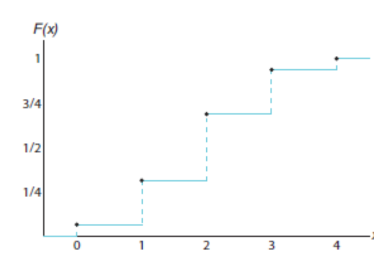

Figure 2: CDF from discrete points. Source: textbook

Let RV X have PMF $f(x)$. The CDF of X is,

$$F(x) = \sum_{t \leq x} f(t)$$

2.2 Continuous Probability Distributions

Let X be a continuous RV. $P(x = 5) = 0$ because the value could be any real number and so $x = 5$ is one of infinite possibilities. [2]

[2] The only case where this isn't true is if there is a Dirac Delta function in the PDF

Probability Density Function (PDF)
$f(x)$ is a PDF of the continuous RV X if

- $f(x) \geq 0$ for each outcome $X = x$
- $\int_{-\infty}^{(} x)dx = 1$
- $\int_a^b f(x)dx = P(a < X < b)$

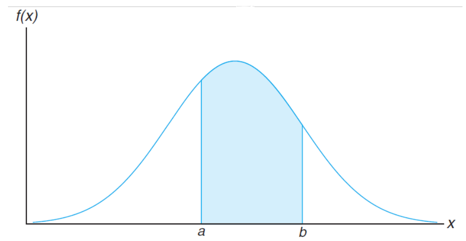

Figure 3: PDF function. Source: textbook

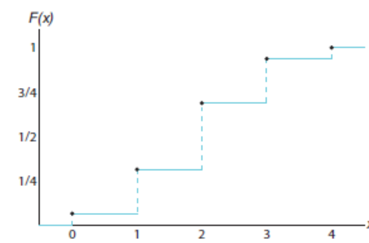

Figure 4: CDF from continuous distribution. Source: textbook

Cumulative Distribution Function (CDF)
Let RV X have PDF $f(x)$. The CDF of X is,

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Some properties:

$$F(x) = P(X \leq x)$$
$$P(a < X < b) = P(b) - P(a)$$
$$F(\text{inf}) = P(X \leq \infty) = \int_{-\infty}^{\infty} f(t)dt = 1$$

2.3 Joint Distributions

Definition: Joint Probability Distribution
If we have two RVs X and Y and we are interested in the simultaneous occurrence of their outcomes (x, y), we can describe the probability distribution through a joint probability distribution function $f(x, y)$.

Discrete: Joint PMF

- $f(x, y) \geq 0 \forall (x, y)$
- $\sum_x \sum_y f(x, y) =$
- $P(X = x, Y = y) = f(x, y)$
- For any region $A \subseteq S$, $P[(X, Y) \in A] = \sum \sum_A f(x, y)$

Continuous: Joint PDF

- $f(x, y) \geq 0 \forall (x, y)$
- $\int_{-\infty}^{\infty} f(x, y)dxdy = 1$
- $P[(X, Y) \in A] = \iint_A f(x, y)dxdy$, for any $A \subseteq A$

2.4 Marginal Distributions

Definition: Marginal Distribution
If we have the joint probability distribution $f(x, y)$ of RVs X, Y and we are interested in the probability distribution of a single RV, we can describe the probability distribution of X and Y through marginal distribution functions $g(x), h(y)$.

[3] **Discrete**

$$g(x) = \sum_y f(x, y)$$
$$h(y) = \sum_x f(x, y)$$

Continuous

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy$$
$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

[3] "Marginal" refers to the marginal totals: the sum of respective columns/rows when the discrete data is presented in a table. We can also think of this as a weighted average of $f(x, y)$ over all possibilities of y or x.

2.5 Conditional Distributions

Definition: Conditional Distribution
The conditional distribution of x given y is,

$$f(x|y) = \frac{f(x, y)}{g(y)}$$

Discrete

$$P(a \leq X \leq b | Y = y) = \sum_{a<x<b} f(x|y)$$

Continuous

$$P(a \leq X \leq b | Y = y) = \int_a^b f(x|y) dx$$

2.6 Statistical Independence

Given RVs X and Y with a joint distribution $f(x, y)$ and marginal distribution $g(x), h(y)$:
X, Y are statistically independent iff $f(x, y) = g(x)h(y), \forall (x, y) \in$ Range.

2.7 Expectation

Definition: Expected Value
Given RV X with distribution $f(x)$, the expected value or mean is given by,

$$\mu = E[X] = \sum_x xf(x)$$
$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

[4] If we have some function $q(X)$ that acts upon the RV X, we get another RV with expected value given by,

$$\mu_{q(x)} = E[q(X)] = \sum_x q(x)f(x)$$
$$\mu_{q(x)} = E[q(X)] = \int_{-\infty}^{\infty} q(x)f(x)dx$$

[4] Mean of a random variables refers to the weighted average, or the arithmetic average of every single option in a discrete case. We are taking into account the relative frequencies of each outcome for the random variable.

If we have RVs X, Y with joint probability distribution $f(x, y)$ and some function $q(X, Y)$, the expected value of RV $q(X, Y)$ is given by,

$$\mu_{q(X,Y)} = E[q(X,Y)] = \sum_x \sum_y q(x,y)f(x,y)$$
$$\mu_{q(X,Y)} = E[q(X,Y)] = \int_{-\infty}^{\infty} q(x,y)f(x,y)dxdy$$

2.8 Variance

Definition: Variance
Let X be an RV with distribution $f(x)$ and mean $\mu = E[X]$. The variance is defined as:

$$\sigma^2 = var(X) = E[(X-\mu)^2] = \begin{cases} \text{Discrete} & \sum_x (x-\mu)^2 f(x) \\ \text{Continuous} & \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \end{cases}$$

We can also express variance of a RV X by

$$\sigma^2 = E(X^2) - \mu^2$$

[5]

[5] We are essentially summing the distance an outcome is from the mean, and multiplying it by the probability that outcome occurs

Definition: Standard Deviation
Standard deviation $\sigma = \sqrt{\sigma^2}$

Definition: Covariance
Let X, Y be RVs with joint distribution $f(x, y)$ and means μ_x and μ_y. The

covariance of X and Y is defined as

$$\sigma_{XY} = cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$
$$= \begin{cases} \text{Discrete} & \sum_x \sum_y (x-\mu_x)(y-\mu_y)f(x,y) \\ \text{Continuous} & \int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)f(x,y)dxdy \end{cases}$$

We can also express covariance by

$$\sigma_{XY} = E[XY] - \mu_x\mu_y$$

If X is positive when Y is positive, $\sigma_{xy} > 0$. If X and Y are inversely related, then $\sigma_{xy} < 0$.

Definition: Correlation Coefficient
The correlation coefficient ρ_{XY} is a normalized covariance given by,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_x}\sigma_y \in [-1,1]$$

2.9 Linear Combinations of Random Variables

The expected value is a linear function. Thus, given RVs X, Y and joint distribution $f(x,y)$,

$$E[aX+Y] = aE[X] + E[Y]$$

Similarly,

$$E[aX+b] = aE(x) + b$$
$$E[q(X,Y) \pm p(X,Y)] = E[g(X,Y)] \pm E[p(X,Y)]$$

If X and Y are independent RVs, then [6]

$$E[XY] = E[X]E[Y]$$

Similarly, we can use linearity with variance to obtain

$$var(aX+bY+c) = a^2var(X) + 2abcov(XY) + b^2var(Y)$$

[6] Recall, $\sigma_{XY} = E[XY] - E[X]E[Y]$. In the independent case, the covariance becomes 0 and this means the variabels are uncorrelated. Note that independent variables are uncorrelated but variables could be uncorrelated and dependent.

3 Common Distribution

Definition: Uniform Distribution
Every element in S has the same probability.

3.1 Binomial Distribution

Definition: Bernoulli Process and Trial
For a sample space of two outcomes (success or failure), a process that consists of n repeated independent trials and has constant probabilities for success and failure is known as a Bernoulli process. Each trial in a Bernoulli

process is known as a Bernoulli trial.

Denote X as the number of successes that occur in n Bernoulli trials. This is known as a binomial RV and its probability distribution is called the binomial distribution. Its values will be denoted using $b(x; n, p)$, where x is the value the RV takes and p is the probability of success for each trial.

The probability of x 1's and $n-x$ 0's in some particular order is $p^x q^{n-x}$, where $q = 1-p$. Thus, the number of ways total to have x 1's and $1-x$ 0's is the above expression multiplied by the number of partitions of n outcomes, or $\binom{n}{x}$. Therefore,

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}$$

Mean of binomial distribution:

$$E[X] = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = np$$

Variance of binomial distribution:
Each trial is independent, so

$$\begin{aligned}\sigma_X^2 = \sum_{k=1}^{n} {\sigma_Y}_k^2 \\ &= \sum_{k=1}^{n} E[Y_k^2] - \mu_{Y\,k}^2 \\ &= \sum_{k=1}^{n} p - p^2 = np(1-p)\end{aligned}$$

3.2 Multinomial Distribution

Multinomial is pretty much the same as binomial except each trial can have more than two outcomes. The notation is then, $f(x_1, x_2, \dots x_k; p_1, p_2, \dots p_k; n)$.
Similar to how the binomial distribution expression was derived, we use partitions to get:

$$f(x_1, \dots x_k; p_1, \dots, p_k, n) = \binom{n}{x_1, \dots, x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

3.3 Hypergeometric Distribution

In binomial and multinomial distributions, we replace whatever was changed from a trial such that the system is identical prior to each trial. In a hypergeometric distribution, we do not replace.

Definition: Hypergeometric Distribution
Assume we have N total objects and we sample n times.
Let K be the number of possible successes out of N.
Let x be the number of successes.
The PMF is defined by:

$$h(x; N, n, K) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

This can be interpreted as (ways to get x successes out of K possible successes)(ways to get $n-x$ failures out of $N-K$ possible failures) / (total possible number of trials).
Valid for $x \geq 0, x \leq n, x \leq K, x \geq n-(N-K)$.

The mean of a hypergeometric distribution is $\mu = n\frac{K}{N}$
The variance is $\sigma^2 = \frac{Kn(N-n)}{N(N-1)}(1-\frac{K}{N})$

3.4 Negative Binomial and Geometric Distributions

Definition: Negative Binomial Distribution
The negative binomial describes the chance that the k^{th} success occurs on the n^{th} trial.
Let RV X be the trial on which the k^{th} success occurs.
The negative binomial distribution is given by:

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

We can derive this from the regular binomial distribution, with the end result being

$$b^*(x; k, p) = p \cdot b(k-1; x-1, p)$$

Definition: Geometric Distribution
The geometric distribution is a special case of the negative binomial with $k=1$. That is, we are looking at RV X with describes the chance that the 1st success on the n^{th} trial.

$$g(x; p) = b^*(x; 1, p) = \binom{x-1}{0} p(1-p)^{x-1} = p(1-p)^{x-1}$$

[7] The mean is given by $\mu = \frac{1}{p}$
The variance is $\sigma^2 = \frac{1-p}{p^2}$

[7] This makes sense because if the first success occurs on trial x, the prior $x-1$ trials must be failures, which mathematically is $p(1-p)^{x-1}$

Example 4
We are playing a game with an opponent that wins 90% of the time (they're just better). How many games should it take until our first victory?
The geometric PMF tells us $g(x; 0.1) = 0.1(0.9)^{x-1}$.
The mean of this PMF is $\frac{1}{0.1} = 10$, so around 10 games until our first win.

3.5 Poisson Distribution

The poisson distribution describes the number of times something happens in a sequence of intervals. [8]

[8] For example, the number of goals in a soccer game or the number of salmon in a river segment.

Definition: Poisson Process
Assume that the number of outcomes in each interval is independent of others (memoryless). In a Poisson Process, the probability of the outcome is proportional to the size of the interval. The number of outcomes in each interval can then be described by a Poisson distribution.

> **Definition: Poisson Distribution**
> Let X be a Poisson RV with PMF:
>
> $$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

The mean and variance is given by $\mu = \sigma^2 = \lambda$. We can interpret this as saying λ is the average number of occurrences per interval.
If r is the rate of occurance and t is the length of the interval, then we can right $\lambda = rt$.
[9]

[9] The limit of the binomial distribution can be shown to be equivalent to the Poisson distribution.

$$\lim_{n\to\infty, p\to 0} b(x;n,p) = p(x;\lambda)$$

3.6 Chi-Squared Distribution

The PDF of the χ^2 distribution is given by,

$$f(x;v) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{x}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Where, v is the degrees of freedom. This is a special case of the gamma distribution with $\beta = 2, \alpha = \frac{v}{2}$.

3.7 Exponential Distribution

The PDF of the exponential distribution is given by,

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x \geq 0 \\ 0, x < 0 \end{cases}$$

This is a special case of the gamma distribution with $\beta = \beta, \alpha = 1$. Since this is still a gamma distribution, we know the mean is β and variance is β^2.
Relationship to the Poisson Distribution The exponential distribution can describe the time taken between events as a complement to the Poisson PMF [10]
The probability of no arrivals in an interval of length t is

$$p(0;rt) = e^{-rt}$$

Let Y be the RV describing the time to first arrival.
Then,

$$P(Y > t) = e^{-rt}$$

[11]
Thus, the CDF of Y is

$$P(Y \leq t) = 1 - e^{-rt} = F(t$$

And it follows that the PDF of Y is

$$f(t) = \frac{d}{dt}F(t) = re^{-rt}$$

$$\text{of arrivals: } p(x;rt) = \frac{e^{-rt}(rt)^x}{x!}$$

$$\text{time to 1st arrival: } f(t) = re^{-rt}$$

[10] Recall: Poisson PMF

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x \in \mathbb{N}$$

$\lambda = rt$, r is the rate of arrivals and t is the length of the interval. Although the events themselves are discrete, the time between is continuous and can be modelled by the exponential distribution.

[11] The point is that, the pro
arrivals during an interval i
the probability of the first a
occurring after the interval

Note that this is the exponential distribution with $r = \frac{1}{\beta}$
Memoryless Property Let X be an exponential distribution with $\beta > 0$.

$$\begin{aligned}
P(X > s + t | X > s) &= \frac{P(X > s + t \cap X > s)}{P(X > s)} \\
&= \frac{P(X > s + t)}{P(X > s)} \\
&= \frac{\int_{s+t}^{\infty} \frac{1}{\beta} e^{\frac{-x}{\beta}} \, dx}{\int_{s}^{\infty} \frac{1}{\beta} e^{\frac{-x}{\beta}} \, dx} \\
&= \frac{e^{\frac{-(s+t)}{\beta}}}{e^{\frac{-s}{\beta}}} \\
&= e^{-\frac{t}{\beta}} \\
&= P(X > t)
\end{aligned}$$

Conclusion is that the prior amount of time spent waiting has no affect on the probability of the wait.

4 Functions of Random Variables

4.1 Functions of Discrete Random Variables

Let X be an RV with PMF $f(x)$ and $Y = u(X)$, where u is bijective. [12] Thus, we can write $X = u^{-1}(Y)$.
The PMF of Y is then given by

$$\begin{aligned}
g(y) &= P(Y = y) \\
&= P(u^{-1}(Y) = u^{-1}(y)) \\
&= P(X = u^{-1}(y)) \\
&= f(u^{-1}(y))
\end{aligned}$$

If X has PMF $f(x)$ and $Y = u(x)$, then $g(y) = f(u^{-1}(x))$.

[12] Bijective means that the function is one-to-one or invertable.

Example 5
X has PMF,

$$f(X) = \begin{cases} \frac{1}{n} & 1 \le x \le n \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$Y = X^2, u(X) = X^2$$

Although $u(x)$ is not invertible in general, it is invertible on the domain of $x \in [1, n]$, the inverse being $u^{-1}(Y) = \sqrt{Y}$ Then,

$$g(y) = \begin{cases} \frac{1}{n} & 1 \le \sqrt{y} \le n \\ 0 & \text{otherwise} \end{cases}$$

Where, $y = 1, 4, 9, \dots n^2$ because X is a discrete RV.

4.2 Functions of Continuous Random Variables

Let X be an RV with PDF $f(x)$ and $Y = u(x)$, where u is bijective. Thus, we can write $X = u^{-1}(Y)$.

The PDF of Y, $g(y)$ cannot be as easily computed as the discrete case. We start with the CDF $G(y)$, [13]

$$\begin{aligned}
G(y) &= P(Y \leq y) \\
&= P(u^{-1}(Y) \leq u^{-1}(y)) \\
&= P(X \leq u^{-1}(y)) \\
&= P(X \leq u^{-1}(y)) \\
&= \int_{\infty}^{u^{-1}(x)} f(x)dx
\end{aligned}$$

$$\begin{aligned}
g(y) &= \frac{d}{dy}G(y) \\
&= f(u^{-1}(y)) \left| \frac{d}{dy} u^{-1}(y) \right|
\end{aligned}$$

[13] We assume u is strictly increasing, which means that the inequality between Y and y also holds between $u^{-1}(Y)$ and $u^{-1}(y)$ Recall that a CDF can either be strictly increasing or decreasing. If we do the calculation with a strictly decreasing, all that changes is the sign of the derivative in the end step so we place an absolute value around it.

4.3 Moment Generating Functions

Definition: Moment Generating Functions (MGF)
The r$^{\text{th}}$ moment of the RV X is,

$$\begin{aligned}
u'_r &= E[X^r] \\
&= \begin{cases} \sum_x x^r f(x) & \text{if discrete} \\ \int_0^{\infty} x^r f(x)dx & \text{if continuous} \end{cases}
\end{aligned}$$

Where, u'_r is the mean u and variance is related by $\sigma^2 = E[X^2] = \mu^2 = \mu'_r - \mu^2$.

Definition: Moment Generating Function of RVs
The MGF of X is given by

$$M_x(t) = E[e^{tx}] \qquad = \begin{cases} \sum_x e^{tx} f(x) & \text{if discrete} \\ \int_{\infty}^{\infty} e^{tx} f(x)dx & \text{if continuous} \end{cases}$$

Observe that the moment generating function can produce all of the moments of X. [14]

$$\begin{aligned}
\frac{d^r}{dt^r} M_x(t)|_{t=0} &= \frac{d^r}{dt^r} \sum_x e^{tx} f(x)|_{t=0} \\
&= \sum_x f(x) \frac{d^r}{dt^r} e^{tx}|_{t=0} \\
&= \sum_x f(x) x^r e^{tx}|_{t=0} \\
&= \sum_x f(x) x^r \\
&= \mu'_r
\end{aligned}$$

[14] Make sure that, when evaluating a function, to substitute after differentiating everything.

Example 6
X is a normal distribution with $n(x; \mu, \sigma)$.

$$
\begin{aligned}
M_x(t) &= \int_{\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2 - 2tx\sigma^2}{2\sigma^2}} dx \\
&= \int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2 - 2(\mu + t\sigma^2)x + \mu^2}{2\sigma^2}} dx
\end{aligned}
$$

Completing the square and doing lots of simplification gives us,

$$
= e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-(\mu+t\sigma^2))^2}{2\sigma^2}} dx
$$

Notice the integrand is just a normal distribution integrated from $-\infty$ to ∞, which equals 1. Thus,

$$
= e^{\mu t + \frac{t^2\sigma^2}{2}}
$$

4.4 Linear Combinations of Random Variables

Let $Y = aX, a \in$. If discrete, then $X = \frac{1}{a}y$. If continuous, then $X = \frac{1}{|a|}f(\frac{1}{a}y)$. Suppose X has MGF $M_x(t)$. Then the MGF of Y is given by,

$$
\begin{aligned}
M_Y(t) &= \int_{-\infty}^{\infty} e^{ty} g(y) dy \\
&= \int_{-\infty}^{\infty} e^{ty} \frac{1}{|a|} f(\frac{y}{a}) dy
\end{aligned}
$$

Let $z = y/a, dy = adz$.

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} e^{az} \frac{1}{|a|} f(z) |a| dz \\
&= M_x(at)
\end{aligned}
$$

Therefore, if X has MGF $M_x(t)$, then $Y = aX$ has $M_Y(t) = M_{ax}(t) = M_x(at)$
Let X have a PMF $f(x)$, Y has a PMF $g(y)$. Let $Z = X + Y$. The PMF of Z is given by,

$$
\begin{aligned}
h(z) &= P(Z = z) \\
&= P(X + Y = z) \\
&= \sum_{w=-\infty}^{\infty} P(X = w) P(Y = z - w) \\
&= \sum_{w} f(w) g(z - w).
\end{aligned}
$$

[15] In the continuous case, we have

[15] note that this is a convolution

$$
h(z) = \int_{-\infty}^{\infty} f(w) g(z - w) dw
$$

Example 7
We have 2 dice, what are the chances of rolling an 8?
Let the PMF for die 1 be $f(x) = \frac{1}{6}$ and that for die 2 be $g(x) = \frac{1}{6}, 1 \leq x \leq 6$.
Let $Z = X + Y$, with PMF $h(z) = \sum_w f(w)g(z - w)$.
Then $h(8) = f(2)g(6) + f(3)g(5) + f(4)g(4) + f(5)g(2) + f(6)g(2) = 5(\frac{1}{6})^2 = \frac{5}{36}$.

X has MGF $M_x(t)$, Y has MGF $M_y(t)$. Let $Z = X + Y$. The MGF of Z is given by

$$\begin{aligned}
M_z(t) &= \sum_z e^{tz} \sum_w f(w)g(z-w) \\
&= \sum_w f(w) \sum_z e^{tz} g(z-w) \\
\text{Let } k = z - w \\
&= \sum_w f(w) \sum_k e^{t(k+w)} g(k) \\
&= \sum_w f(w)e^{tw} \sum_k e^{tk} g(k) \\
&= M_x(t)M_y(t)
\end{aligned}$$

5 Sampling

A sample is a subset of a population. Each observation, x_i is the realization of a RV X_i.
The joint distribution $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$. Thus, the X_i's are independent.
A statistic is a function of the X_i's. A statistic is biased if it is consistently over or under estimates.
The following are some sample statistics:

Sample Mean
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample Median
If we arrange data in increasing order, $x_1, x_2, \ldots x_n$:

$$x_m = \begin{cases} \frac{x_{n/2}+x_{n/2+1}}{2}, & \text{if n is even} \\ x_{(n+1)/2}, & \text{if n is odd} \end{cases}$$

$$X_m = \begin{cases} \frac{X_{n/2}+X_{n/2+1}}{2}, & \text{if n is even} \\ X_{(n+1)/2}, & \text{if n is odd} \end{cases}$$

Mode
Most frequently occurring value.

Sample Variance

$$S^2 = \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample standard deviation: $s = \sqrt{S^2}$

Unbiased Statistics Each x_i is a realization of X with PDF $f(x_i)$, mean μ, and variance σ^2.
The expectation of sample mean:

$$\begin{aligned} E[\bar{X}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n} n\mu = \mu \end{aligned}$$

$\bar{X}$ is an unbiased estimate of μ.
For variance, we want $E[S]^2 = \sigma^2$

$$\begin{aligned} \sigma^2 &= E[X_i^2] - \mu^2 \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 + \bar{X}^2 - 2X_i\bar{X} \\ &= \frac{1}{n-1}(n\bar{X}^2 - 2n\bar{X}^2 + \sum_{i=1}^{n} X_i^2) \\ &= \frac{1}{n-1}(-n\bar{X}^2 + \sum_{i=1}^{n} X_i^2) \\ E[S^2] &= \frac{1}{n-1}(-nE[\bar{X}^2] + \sum_{i=1}^{n} E[X_i^2]) \\ &= \frac{1}{n-1}(-n(E[\bar{X}]^2 + var[\bar{X}]) + \sum_{i-1}^{n} \sigma^2 + \mu^2) \\ &= \frac{1}{n-1}(-n(\mu^2 + \frac{1}{n}\sum_{i=1}^{n} var[X_i]) + n(\sigma^2 + \mu^2)) \\ &= \frac{1}{n-1}(-n(\mu^2 + \frac{1}{n^2} n\sigma^2) + n(\sigma^2 + \mu^2)) \\ &= \frac{1}{n-1}(n-1)\sigma^2 \\ &= \sigma^2 \end{aligned}$$

Therefore,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an unbiased estimator. [16]

The distribution of a statistic is a **sampling distribution**. [17]

Useful Facts about Statistics

- If X_1 and X_2 are normal with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Then $X_1 + X_2$ is normal with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.
- $\frac{1}{n}X_1$ is a RV with mean $\frac{\mu_1}{n}$ and variance $\frac{\sigma_1^2}{n^2}$.

[16] We use n-1 because we need to account for using the sample mean instead of the true mean.
[17] For example, in $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, each X_i is a sample with some distribution so $\bar{X}$ is a sampling distribution.

- Thus, if $X_1, \dots, X_n$ are normal, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normal with mean μ and variance $\frac{\sigma^2}{n}$

Theorem: Central Limit Theorem
Assume we have a sample $X_1, X_2, \dots, X_n$, in which X_i's are independent identically distributed (IID).
Let the mean be μ and let the finite variance be σ^2.
Let $X_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Z_n = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}$.
The central limit states that: as $n \to \infty$, the distribution of Z_n approaches $n(z; 0, 1)$.

Given mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$Z_n = \frac{n - \mu}{\sigma}\sqrt{n} = (\bar{X} - \mu)\frac{\sqrt{n}\sigma}{}$$

Then by the CLT: the distribution of Z_n as $n \to \infty$ is $n(z; 0, 1)$.
The standard deviation of $\bar{X}_n$ is approximately $\frac{\sigma}{\sqrt{n}}$.

Example 8
Consider a runner who can run at 4 min/mile with $\sigma = 5$. What is the chance the average of the next 20 races is below $3:58$? We want $P(\bar{X}_{20} \leq 238s)$:

$$= P(\frac{\bar{X} - 240}{5}\sqrt{20} \leq \frac{238 - 240}{5}\sqrt{20})$$

The first term, Z_{20}, approximately has a PDF $n(z; 0, 1)$.

$$\begin{aligned} &= P(Z_{20} \leq -1.8) \\ &\approx \int_{-\infty}^{-1.8} n(z; 0, 1) dz \\ &= \Phi(-1.8) = 0.036 \end{aligned}$$

5.1 Chi-Squared Distribution

The distribution of the sample variance is given by the χ^2 distribution:

$$f(x; v) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{v/2-1} e^{-x/2}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Assuming we have samples $X_1, X_2, \dots X_n$, each being a normal distribution with variance σ^2, the χ^2 statistic is given by

$$\chi^2 = \frac{n-1}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

This has the distribution $f(x; v)$ with $v = n - 1$, where v is known as the degrees of freedom.

What if the distribution was instead based on the true mean?

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

The distribution would instead by $f(x; v), v = n$.
v represents the amount of information we have, by knowing the true mean we go from $n-1$ to n.

5.2 T-Test Statistic

Prior, we had a known variance. Consider sample $X_1, X_2, \dots X_n$, where X_i is normal with mean μ but σ^2 is unknown. Let

$$T = \frac{\bar{X} - \mu}{s}\sqrt{n}$$

This has a distribution (t-distribution)

$$h(t; v) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{\pi v}}(1 + \frac{t^2}{v})^{\frac{-(v+1)}{2}}$$

Where T has a t-distribution with $v = n - 1$.

6 Appendix A

6.1 Approaches to Combinatorics Problems

Arrangements
The number of arrangements of k for a set of n elements is given by n^k.

Examples:

- Thirteen pairs of football teams will play matches today. The possible predictions for every match are victory for the host team, loss for the host team, and a draw. How many different predictions are there?
  There are 3 different options for 13 events so the total number of possible predictions is 3^{13}

Arrangements Without Replacement/Repetition
The number of arrangements of k without replacement for a set of n elements is given by $n(n-1)\dots(n-k+1)$.

Examples:

- How many ways can we elect a president, vice president, and treasurer if there are 25 member of a club? Here, $n = 25$ and $k = 3$ so no. of ways $= (25)(24)(23)$

Permutations
idk
Identical Items: If we have repetitions in a set, we divide by the number of repeated elements factorial (removing the permutations of repeated elements).

Examples:

- Number of unique strings that can be permuted from "ATLANTIC": we have 8 letters, but 'A' and 'T' are repeated twice. Thus, permuations $= \frac{8!}{2!2!}$.

- Same problem as above, but the 'A's and 'T's must not be adjacent to each other. We consider the complement: how many permutations are there such that the 'A's are adjacent and 'T's are adjacent. Since they are identical, we can group them into a single term, 'AA' and 'TT'. Then we have 6 letters, so the complement has 6! ways. The number of ways for no-adjacent repeats is then $8! - 6!$.

- 10 coin flips to get 4 heads: reword to get "number of unique arrangements of "HHHHTTTTTT" (4 heads, 6 tails). This is the same problem as earlier, ways $= \frac{10!}{4!6!}$

- Given the set of 16 letters $\{a, b, c, ...p\}$, how many permutations are there where 'a' and 'g' are separated by exactly 4 letters: there are 11 possible positions for "a . . . . g" and we can also have "g . . . . a", the other 14 letters can be arranged any way we like. Thus, of ways $= (11)(2)(14!)]$

Combinations
idk either
Examples:

- 12 boys and 8 girls in a class and we need to choose three of each. How many ways can we do this?
  choose boys $= \binom{12}{3}$, choose girls $= \binom{8}{3}$, of ways $= \binom{12}{3}\binom{8}{3}$

- Number of ways to get two pairs in a 6-card hand from a standard deck. First we choose a card out of 13 options, then choose 2 from the 4 suits: $\binom{13}{1}\binom{4}{2}$. Then we repeat for the remaining 12 options, choosing 2 from the 4 suits $\binom{12}{1}\binom{4}{2}$. Now we have 2 pairs and choose the remaining 2 cards from the 11 and choose their colours $\binom{11}{2}\binom{4}{1}\binom{4}{1}$. The number of ways then is the product of all of the earlier terms.

- Number of ways to get a straight (increasing sequence) in a 5-card hand from a standard deck. We don't care what order we get the cards in, so there's only 9 possible straights starting with A-5 and ending with 9-K. We have to choose a colour from each so the number of ways is $9\binom{4}{1}^5$

- Number of ways to get a flush (all same suit, but not increasing sequence). We first choose a suit $\binom{4}{1}$. Then we choose 5 cards from it $\binom{13}{5}$, but subtract the number of ways where we get a straight 9. Thus, the number of ways to get a flush is $\binom{4}{1}(\binom{13}{5} - 9)$.

To be continued...

6.2 Notes on Probability Problems

For a system with 4 components $A - D$, but redundant / parallel subsystems B and C, the probability of success can be found by:

$$P = P(A)[1 - (1 - P(B))(1 - P(C))]P(D)$$

The probability of the redundant section succeeding is $1-$(probability of both failing).

This idea can be similarly extended to another problem: if we are interested in the probability of something occurring more than n times, it is typically easier to find

$1-$ (probability of it happening $\leq n$ times) because we don't involve infinity.
If

$$f(x) = e^{-6}\frac{6^x}{x!}, x = 0, 1, 2, ...$$
$$P(X < 8) = 1 - \sum_{x=0}^{8} e^{-6}\frac{6^x}{x!}$$

For a joint distribution, $P(X < Y)$ can be found setting the upper bound of the x integral to y:

$$P(X, Y) = \int_a^b \int_c^y f(x, y)dxdy$$

7 Appendix B

7.1 Applications in Renewable Energy

- Popular forms of renewable energy, especially wind and solar, suffer from intermittency
- This can be costly, since an alternative source of energy must be used to generate energy in times of intermittent outage

Let a given renewable energy source produce p units of energy, with distribution $f(p)$. Let the base amount paid be λp (proportional to produced energy) and the forecast be denoted $\widehat{p}$. Then, let the payment be instead: [18]

$$\lambda p - \mu^-(\widehat{p} - p)^+ - \mu^+(p - \widehat{p})^+$$

Where, μ^+, μ^- are constants. So the payment is basically the original payment minus a fee due to underproduction and a fee due to overproduction. [19]
We want a forecast $\widehat{p}$ that maximizes the payment. Since we cannot predict the future, we want to maximize our expected profit:

$$\begin{aligned} J &= E[\lambda p - \mu^-(\widehat{p} - p)^+ - \mu^+(p - \widehat{p})^+] \\ &= \lambda \int_{-\infty}^{\infty} pf(p)dp - \mu^- \int_{-\infty}^{\infty} (\widehat{p} - p)^+ f(p)dp - \mu^+ \int_{-\infty}^{\infty} (p - \widehat{p})^+ f(p)dp \end{aligned}$$

[20] We want to maximize this expression, so finding the critical points by differentiating and a bunch of math using Leibniz's integral rule:

$$\frac{dJ(\widehat{p})}{d\widehat{p}} = 0$$
$$\frac{d}{d\widehat{p}}\lambda \int_{-\infty}^{\infty} pf(p)dp = 0$$
$$\frac{d}{d\widehat{p}}\mu^- \int_{-\infty}^{\infty} (\widehat{p} - p)^+ f(p)dp = \int_{-\infty}^{\widehat{p}} f(p)dp = F(\widehat{p})$$
$$\frac{d}{d\widehat{p}}\mu^+ \int_{-\infty}^{\infty} (p - \widehat{p})^+ f(p)dp = \int_{-\infty}^{\widehat{p}} -f(p)dp = -(1 - F(\widehat{p}))$$
$$\frac{dJ(\widehat{p})}{d\widehat{p}} = -\mu^- F(\widehat{p}) + \mu^+(1 - F(\widehat{p})) = 0$$
$$\therefore F(\widehat{p}) = \frac{\mu^+}{\mu^+ + \mu^-}$$

[18] We denote $(x)^+ = \max(x, \sigma)$.

[19] e.g. if $p > \widehat{p}$ (overproduction), the $\mu^-(\widehat{p} - p)^+$ term goes to 0 and overproduction fee is non-zero.

[20] Since the underproduction/overproduction fees are bounded, we can set the integrals bounds to $\int_{-\infty}^{\widehat{p}}$ and $\int_{\widehat{p}}^{\infty}$ respectively. Note that later on we can also rewrite an integral with these bounds as a CDF.

We know that a PDF of wind is invertible so, $\widehat{p}^* = F^{-1}(\frac{\mu^+}{\mu^+ + \mu^-})$. If $\mu^+ \gg \mu^-$: $\widehat{p}^* \to \infty$ and $\mu^- \gg \mu^+ : \widehat{p}^* \to 0$. Note that if $\mu^+ = \mu^-, \widehat{p} = \frac{1}{2}$ is the median.